Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Feature and Lyrics

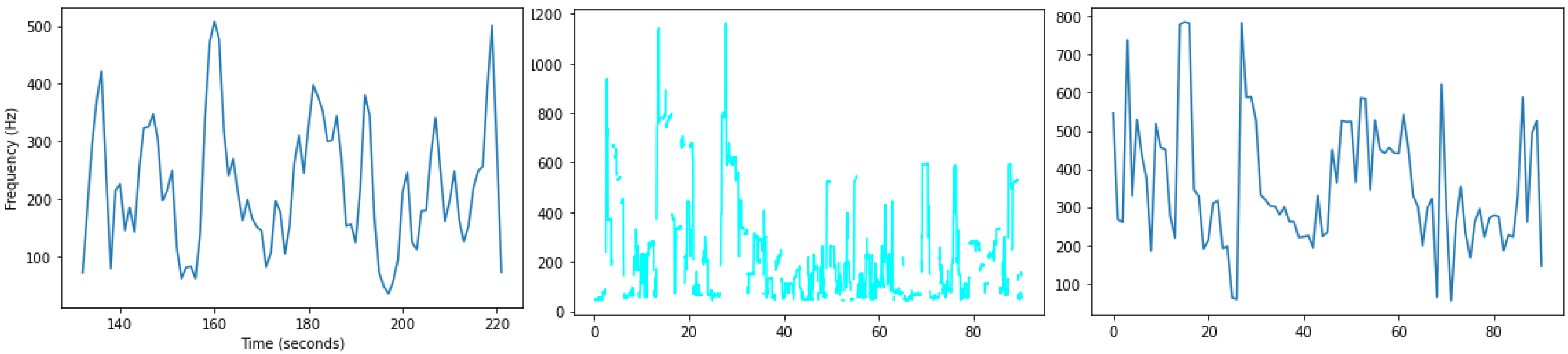
Krols, Tibor, Yana Nikolova, and Ninell Oldenburg (University of Copenhagen). "Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Features and Lyrics." *arXiv preprint arXiv:2302.13321* (2023)

경영과학연구실 이태헌
2023.03.15

Do you know what logo is this?

F0 Estimator 비교

- DEAM Groundtruth 값과 PYIN, CREPE 알고리즘 F0 값 비교



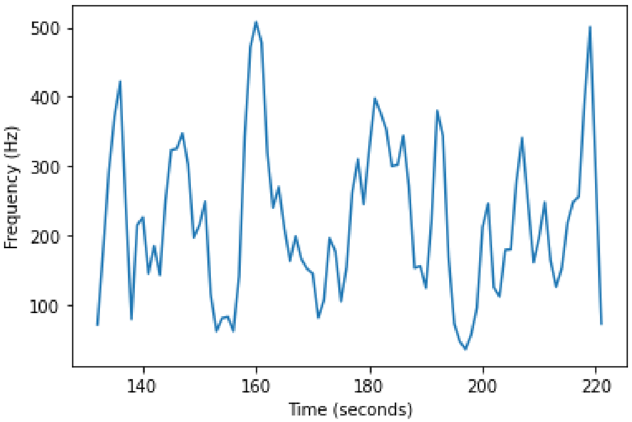
Why are high-level features necessary?

1. Music is one of the most complex forms of art created by humans
2. Music provides a highly subjective experience to people

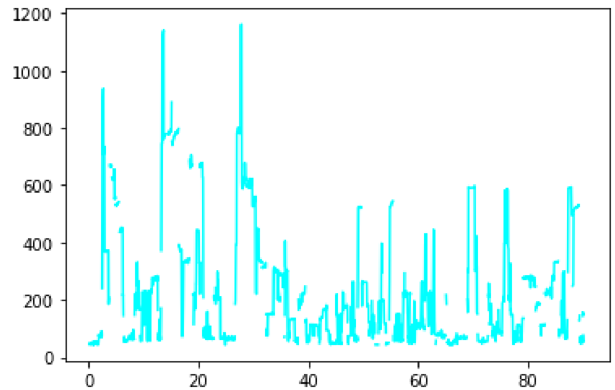
- A single song is composed of thousands of low-level features, and each feature interacts with each other to create the unique characteristics of a song

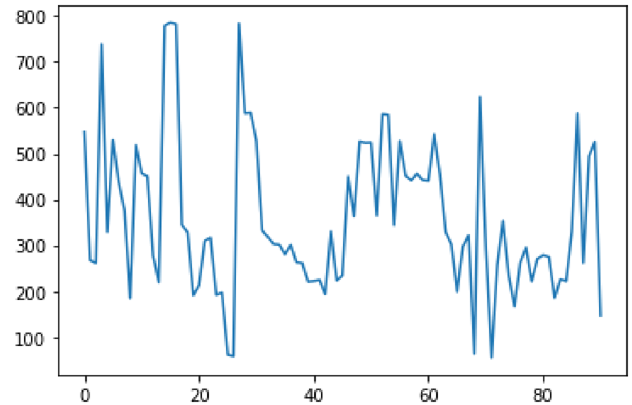
- High-level features are typically obtained by combining and analyzing the characteristics of low-level features extracted from music data

Combination of low-level features (Frequency, pitch, Chord) → **High-level features**



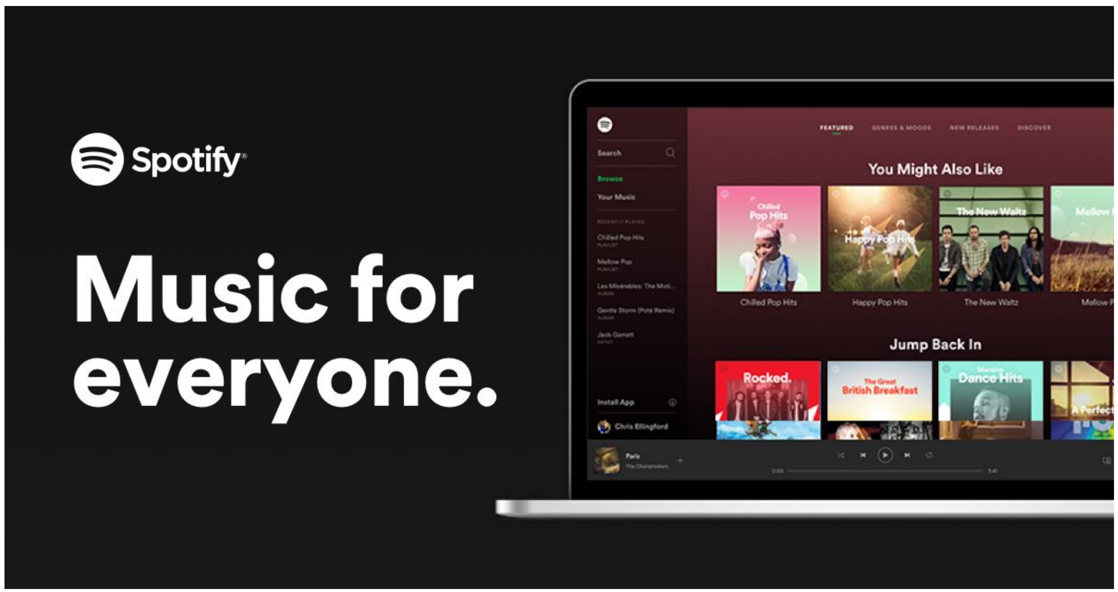
Frequency (Hz)
Time (seconds)

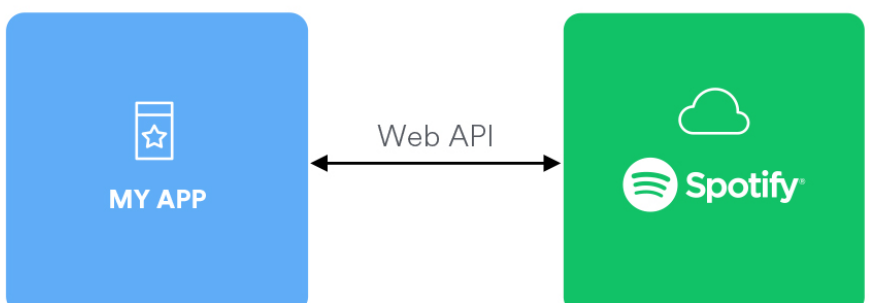




Spotify

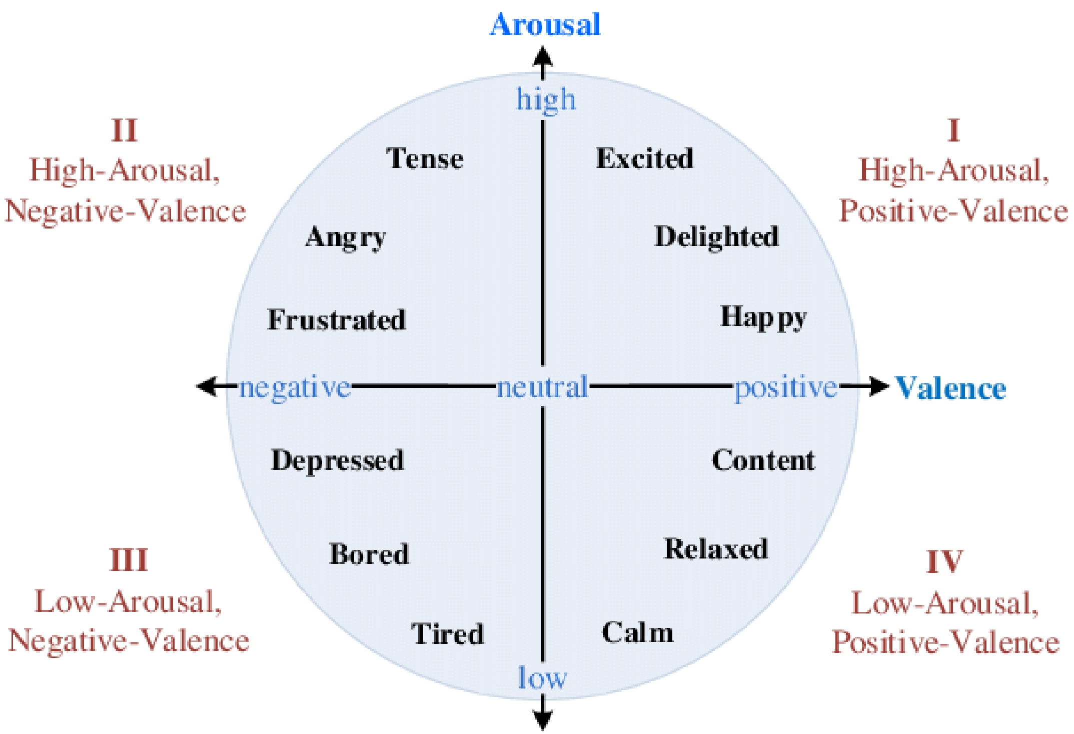
- Spotify is one of the most popular music streaming services in the world, with over 70 million users worldwide





Valence-Arousal space

- Valence-Arousal space is a 2-dimensional coordinate system used to represent emotions
- Valence represents the degree of positive/negative emotion, while Arousal represents the degree of activity/calmness of the emotion
- They are measured on a scale of -1 to 1, depending on the degree



Russell, A circumplex model of affect (1980)

Spotify open API feature

- Used features and description taken from the Spotfiy documentation

Feature	Description
Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic
Danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity
Energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity
Instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context
Key	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation
Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live
Loudness	The overall loudness of a track in decibels (dB)
Mode	Indicates the modality (major or minor) of a track
Speechiness	Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry)
Tempo	The overall estimated tempo of a track in beats per minute (BPM)
Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track

Why MER(Music Emotion Recognition) is difficult

- Lack of clear benchmark data and measurement metrics for results

Speech Emotion Recognition

72 papers with code • 13 benchmarks • 14 datasets

Categorical speech emotion recognition. Emotion categories: Happy (+ excitement), Sad, Neutral, Angry Modality: Speech Only

For multimodal emotion recognition, please upload your result to Multimodal Emotion Recognition on IEMOCAP

Benchmarks

Add a Result

These leaderboards are used to track progress in Speech Emotion Recognition

Trend	Dataset	Best Model	Paper	Code	Compare
	IEMOCAP	DANN			See all
	CREMA-D	SepTr + LeRaC			See all
	RAVDESS	TIM-Net			See all

Rank	Model	WA	UA	F1	Accuracy	Macro Recall	Paper
1	DANN	0.827	-	-			Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition
2	TAP	0.81					Speaker Normalization for Self-supervised Speech Emotion Recognition
3	SYSCOMB: BLSTMATT with CSA	0.805	0.74	-			Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition
4	Partially Fine-tuned HuBERT Large	0.796					A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding
5	LSTM+FC	0.755	-	-			Speech Emotion Recognition Using Speech Feature and Word Embedding

Music Emotion Recognition

5 papers with code • 0 benchmarks • 2 datasets

This task has no description! Would you like to contribute one?

Benchmarks

Add a Result

These leaderboards are used to track progress in Music Emotion Recognition

No evaluation results yet. Help compare methods by submitting evaluation metrics.

Datasets

RAVDESS VGMIDI

Most implemented papers

Most implemented Social Latest No code

Search for a paper, author or keyword

Tracing Back Music Emotion Predictions to Sound Sources and Intuitive Perceptual Qualities

2

CPJKU/audioLIME • 14 Jun 2021

In previous work, we have shown how to derive explanations of model predictions in terms of spectrogram image segments that connect to the high-level emotion prediction via a layer of easily interpretable perceptual features.

Paper Code

Music Mood Detection Based On Audio And Lyrics With Deep Neural Net

1

Dohppak/Music-Emotion-Recognition-Classification • PyTorch • 19 Sep 2018

We consider the task of multimodal music mood prediction based on the audio

Paper Code

Related works

MER as Regression Task

- *Yang, Yi-Hsuan, et al. "A regression approach to music emotion recognition." IEEE Transactions on audio, speech, and language processing 16.2 (2008): 448-457.*
- *Vatolkin, Igor, and Anil Nagathil. "Evaluation of audio feature groups for the prediction of arousal and valence in music." Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement (2019): 305-326.*

Lyrics as Prediction Metric

- *Han, Donghong, et al. "A survey of music emotion recognition." Frontiers of Computer Science 16.6 (2022): 166335*
- *Hu, Xiao, Kahyun Choi, and J. Stephen Downie. "A framework for evaluating multimodal music mood classification." Journal of the Association for Information Science and Technology 68.2 (2017): 273-285.*

Higher-level features

- *Panda, Renato, et al. "How Does the Spotify API Compare to the Music Emotion Recognition State-of-the-Art?." 18th Sound and Music Computing Conference (SMC 2021)*
- *Vatolkin, Igor, and Anil Nagathil. "Evaluation of audio feature groups for the prediction of arousal and valence in music." Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement (2019)*

Problem statement & Key idea

Problem statement

This paper aims to address the problem of emotion recognition in music

Key idea

1. This paper uses a multi-modal approach

   - Audio feature : High-Level audio feature (Spotify open API)
   - Lyrics feature : To represent the lyrical information, they created three types of features (Sentiment information, TF-IDF features, ANEW features)

2. This paper combines tag values from DMDD, LastFM, ANEW, and Spotify data

* DMDD : Deezer Mood Detection Dataset
* ANEW : Affective Norms for English Words

Data

- The DMDD, ANEW, and Spotify data were combined and used, involving three stages of preprocessing

1. DMDD (Deezer Mood Detection Dataset)

- Which holds VA scores for 18,644 songs and is based on the Million Song Dataset as well as tags from LastFM that are related to mood (V,A range is 1-9)

E.g. Music – (V : 5, A : 3, sad, tired)

2. VA scores were obtained by applying an extended ANEW (Affective Norms for English Words) dataset

- The dataset is used for studying the relationship between words and emotions. It includes around 14000 English words with emotion weights ranging from 1 to 9
- Measuring three emotional dimensions of words: Valence, Arousal, and Dominance
- With 14,000 words and their respective VA scores to the tags from LastFM

E.g. Music – (sad = V:8, A: 3, tired = V:5, A:1)

3. High-level features for all available songs from the DMDD via the Spotify

- Spotify's valence annotation is derived differently from our ground-truth valence, avoiding circularity and is also used as a predictive feature for emotion in Panda et al.(2021)

Ground truth Valence ≠ Sptofiy Valence

Panda, Renato, et al. "How Does the Spotify API Compare to the Music Emotion Recognition State-of-the-Art?." 18th Sound and Music Computing Conference (SMC 2021)

Extracting Lyrics Features

- Represent the lyrical information, this paper create three types of features

1. Sentiment information

- Consisting of positive, negative, neutral and compound scores was obtained with VADER(Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis

2. TF-IDF(Term Frequency-Inverse Document Frequency) features

- TF-IDF stands for "Term Frequency-Inverse Document Frequency," and it is a method of evaluating how important a specific word is within a document

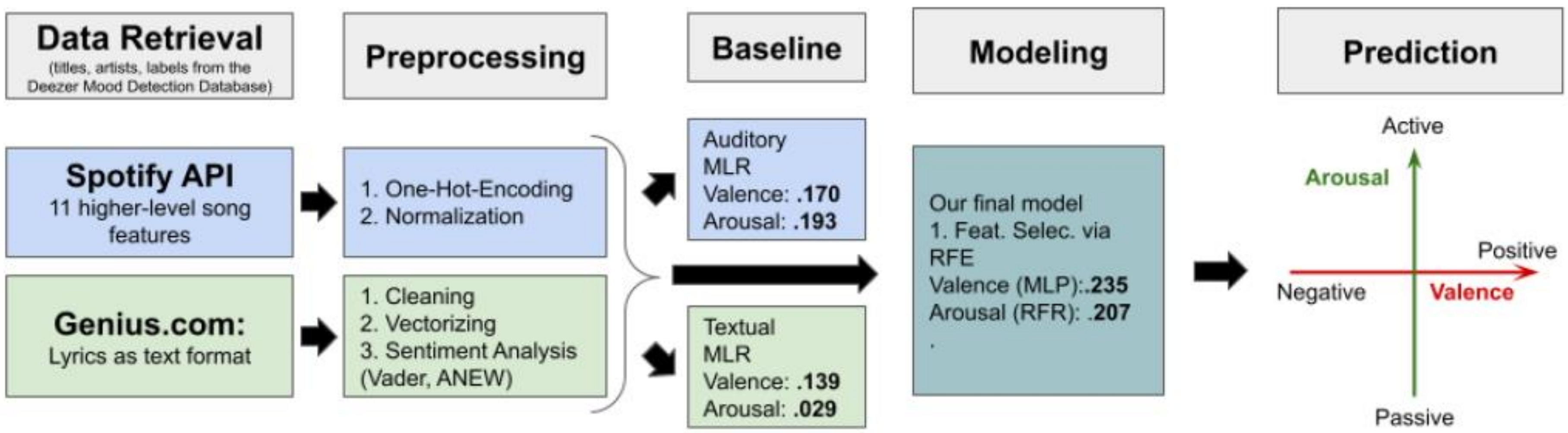
$$TF = \left(\frac{\text{Number of times keyword is found in document}}{\text{Number of words in document}} \right) \quad IDF = \log \left(\frac{\text{Number of documents}}{\text{Number of documents containing the keyword}} \right)$$

3. ANEW features

- They generated two count vectors for each pre-processed lyric text and multiplied the counts by the respective VA scores

Model process

- Audio data : Spotify API
- Lyrics data : Genius.com (crawling)



Model

- MLR, RFR, SVR, MLP

MLR (Multiple Linear Regression)

- A statistical technique for modeling the linear relationship between a dependent variable and one or more independent variables

RFR (Random Forest Regression)

- One of the machine learning techniques for regression analysis. RFR is an ensemble method based on decision trees, which learns multiple decision trees to predict results

SVR (Support Vector Regression)

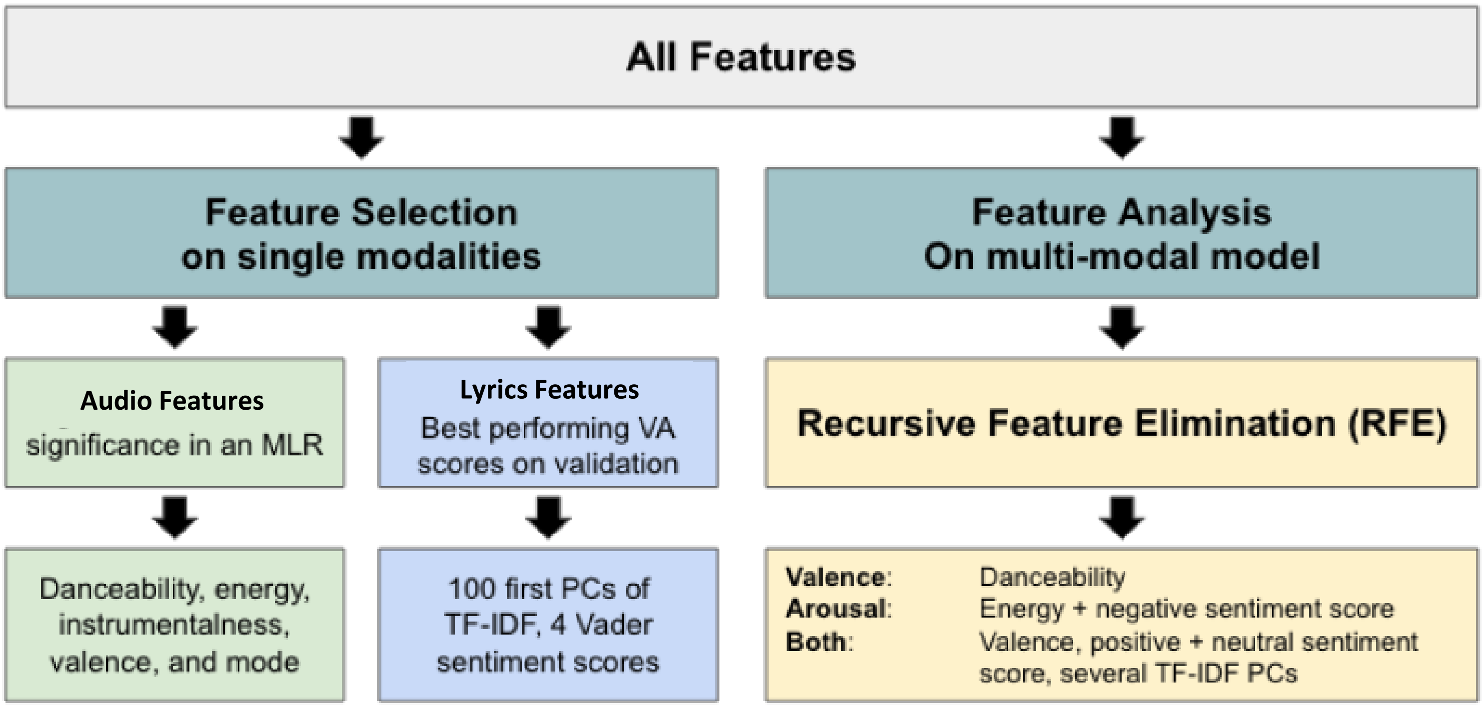
- A machine learning technique for regression analysis. It is a derived algorithm from SVM. SVR performs regression analysis by mapping the data features to a higher-dimensional space and finding the optimal decision boundary (or hyperplane) for regression

MLP (Multi-Layer Perceptron)

- A type of artificial neural network that uses multiple hidden layers to learn complex nonlinear models

Feature selection

- Feature selection



Recursive feature elimination (RFE)
- One of the feature selection techniques used in machine learning. It is a method of iteratively training a model and removing features in order to find the most useful features from a given dataset

Model Results

- R^2 test scores for all uni and multi-modal models based on selected feature subsets

1. all features_A
{Acousticness, Danceability, Energy, Instrumentalness, Key, Liveness, Loudness, Mode, Speechiness, Tempo, Valence}

2. selected_A
{Danceability, Energy, Instrumentalness, Valence, Mode}

3. all features_L
{ANEW scores, TF-IDF, 4 Vader sentiment scores}

4. selected_L
{TF-IDF, 4 Vader sentiment scores}

Mode	Model	Valence	Arousal
Audio	MLR	0.170	0.193
	RFR	0.171	**0.204**
	SVR	0.165	0.203
	MLP	**0.176**	0.203
Lyrics	MLR	**0.139**	**0.029**
	RFR	0.121	0.027
	SVR	0.042	-0.074
	MLP	0.117	0.020
Multi-modal	MLR	**0.236**	0.190
	RFR	0.224	**0.207**
	SVR	0.208	0.154
	MLP	0.235	0.196

Feature Analysis

- p-values of coefficients in MLR
- Valence has 7 significant predictors
- Arousal has 6 significant predictors

Feature	Valence	Arousal
Constant	**-1.6885***	**-0.9836***
Danceability	**0.6915***	**-0.3266***
Energy	**0.6378***	**1.4254***
Loudness	-0.0091	-0.0073
Speechiness	-0.1101	**0.3952***
Acousticness	**0.1649***	0.0207
Instrumentalness	0.0929	**-0.3278***
Liveness	**0.1916***	0.0207
Valence	**1.0901***	**0.5158***
Tempo	0.0005	0.0004
Mode	**0.0977***	**0.1272***
Compound sentiment	**0.2275***	-0.0051

coefficients for MLR. *significant with p <0.05

Constant, Danceability, Energy, Valence, Mode

Selected Features Performance

- Compares VA scores of the MLP with all features vs. selected features for each modality

1. all features_A
{Acousticness, Danceability, Energy, Instrumentalness, Key, Liveness, Loudness, Mode, Speechiness, Tempo, Valence}

2. selected_A
{Danceability, Energy, Instrumentalness, Valence, Mode}

3. all features_L
{ANEW scores, TF-IDF, 4 Vader sentiment scores}

4. selected_L
{TF-IDF, 4 Vader sentiment scores}

	Feature set	Valence	Arousal
Audio	all_features$_A$	0.163	0.193
	selected$_A$	**0.176**	**0.203**
Lyrics	all_features$_L$	0.091	0.009
	selected$_L$	**0.117**	0.019
Multi	all_features$_A$ + all_features$_L$	0.230	0.193
	selected$_A$ + selected$_L$	**0.235**	**0.196**

Comparison of MLP R^2 scores for different feature subsets

Conclusion & Future Directions

Conclusion

- Both uni-modal lyrics features an uni-modal audio features reasonably predict valence, although a multi-modal approach outperforms either modality individually

- Predicting arousal is hard to do with lyrics features, since audio features alone perform almost as well as the multi-modal approach

Future Directions

- Early Feature Fusion -> Late Feature Fusion
- Deep Learning as State-of-the Art
- Vague Annotation Standard

Q & A